Published online 18 November 2011

Nucleic Acids Research, 2012, Vol. 40, No. 2 ***e11***
doi:10.1093/nar/gkr955

i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets

Sebastian Proost[1,2], Jan Fostier[3], Dieter De Witte[3], Bart Dhoedt[3], Piet Demeester[3], Yves Van de Peer[1,2,*] and Klaas Vandepoele[1,2]

[1]Department of Plant Systems Biology, VIB, [2]Department of Plant Biotechnology and [3]Ghent University, B-9050 Ghent, Belgium

Received July 29, 2011; Revised September 29, 2011; Accepted October 13, 2011

ABSTRACT

Comparative genomics is a powerful means to gain insight into the evolutionary processes that shape the genomes of related species. As the number of sequenced genomes increases, the development of software to perform accurate cross-species analyses becomes indispensable. However, many implementations that have the ability to compare multiple genomes exhibit unfavorable computational and memory requirements, limiting the number of genomes that can be analyzed in one run. Here, we present a software package to unveil genomic homology based on the identification of conservation of gene content and gene order (collinearity), i-ADHoRe 3.0, and its application to eukaryotic genomes. The use of efficient algorithms and support for parallel computing enable the analysis of large-scale data sets. Unlike other tools, i-ADHoRe can process the Ensembl data set, containing 49 species, in 1 h. Furthermore, the profile search is more sensitive to detect degenerate genomic homology than chaining pairwise collinearity information based on transitive homology. From ultra-conserved collinear regions between mammals and birds, by integrating coexpression information and protein–protein interactions, we identified more than 400 regions in the human genome showing significant functional coherence. The different algorithmical improvements ensure that i-ADHoRe 3.0 will remain a powerful tool to study genome evolution.

INTRODUCTION

During their evolution, genomes have been altered at various levels. At the smallest scale, point mutations and small insertions and deletions (1) affect only a few nucleotides. Larger modifications include duplication, deletion, translocation or inversion of a single gene or genomic segment (2). At the largest scale, the entire genome can be doubled via genome duplication or merging (3–5). Identification of these structural rearrangements provides insight into how genomes have evolved and diverged over time. It is therefore of crucial importance to correctly determine chromosomal regions that are homologous (i.e. derived from a common ancestor), either *within* a genome, or *between* genomes of related species. Genomic homology can be inferred from collinearity, namely the conservation of both gene content and gene order. Synteny, though initially defined as 'the property of being located on the same chromosome' (6), is often used to indicate the conservation of gene content but not necessarily gene order (7). Like collinearity, synteny also points to homology between different genomic regions based on a number of shared genes (8,9).

Detection of collinear regions between the genomes of related species allows for the identification of chromosomal fusions and fissions, along with inverted or translocated regions. Additionally, gene loss and gain can be efficiently estimated, and cross-species genome analysis provides a framework for transferring gene annotation and biological information to newly sequenced genomes. Finally, orthologous intergenic sequences derived from collinear regions can be screened for conserved non-coding regions as a way to detect regulatory motifs and to identify various types of RNA genes (10). As both gene loss and different types of rearrangements accumulate over time, the resulting genome erosion gradually reduces the degree of collinearity between species. Therefore, gene order preserved over a large phylogenetic distance can imply a biological constraint (11).

Collinear regions within a genome can also hint at the occurrence of one or more rounds of whole-genome duplications (WGDs) (9,12). Based on within-genome collinearity, the loss of gene duplicates created during a WGD

*To whom correspondence should be addressed. Tel: +32 9 331 3807; Fax: +32 9 331 3809; Email: yves.vandepeer@psb.vib-ugent.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

can be estimated (13–15), whereas the functions of genes retained in duplicate can be linked to lineage or species-specific adaptations, including specific pathways and biological processes. WGDs appear to have played a crucial role in the evolution of all major eukaryotic lineages and, particularly in plants, they are often associated with key events during evolution including fast adaptive radiation (4,16) and survival of mass extinction events (17). Additionally, gene family expansions critical for the pome fruit development in apple (*Malus domestica*) (18) have been linked to a recent WGD, whereas expansions in genes producing aromatic compounds have been observed in grapevine (*Vitis vinifera*) (19). Although remnants of several recent WGDs are abundant in the plant kingdom, WGDs in land vertebrates and fishes are seemingly much older (20,21). In vertebrates, the complex body plan is often attributed to the duplication of developmental genes during two WGDs 450 million years ago (Mya) (21). The first traces of a WGD have been unveiled in *Saccharomyces cerevisiae* based on comparative approaches (22). Additional proof for the WGD in brewer's yeast has been provided later by comparison with the genome of an unduplicated outgroup species, *Kluyveromyces waltii* (23). The more complex carbohydrate metabolism of *S. cerevisiae* and other post-duplication yeast species is probably a direct consequence of this duplication (24). Therefore, the discovery of large-scale duplications, through the study of collinear regions, has provided a remarkably detailed view on the genomic evolution and adaptation of various species.

Here, we focus on the accurate detection of homologous chromosomal segments both within and between the genomes of related species. Specifically, sensitive and accurate algorithms are needed for the identification and evolutionary analysis of duplicated regions that have undergone massive gene loss. Several tools, by means of various approaches, have recently been proposed (Supplementary Table S1). Whereas most tools only perform pairwise comparisons, the iterative Automatic Detection of Homologous Regions (i-ADHoRe) (25) was one of the first that simultaneously analyzed genomes of multiple species and allowed for the detection of highly diverged collinear regions. On the one hand, i-ADHoRe has been used in several genome projects to uncover the remnants of large-scale duplications [e.g. apple (*M. domestica*) (18), soybean (*Glycine max*) (26), *Arabidopsis lyrata* (27) and black cottonwood (*Populus trichocarpa*) (28)], and, on the other hand, to detect inter-species collinearity in yeasts (29) and Archaea (30). In contrast to tools that infer genomic homology through a multiple sequence alignment of complete genomic DNA sequences (31–34), i-ADHoRe detects genomic homology through the identification of gene collinearity and/or synteny. The core feature of i-ADHoRe 3.0, which is based on a new alignment algorithm (35) and improved statistical evaluation, is the ability to handle large numbers of genomes. Due to the further optimization of many algorithmic steps, the current version of i-ADHoRe 3.0 is roughly 30 times faster than the previous version. In addition, i-ADHoRe 3.0 can now take advantage of a parallel computing platform, reducing the runtime even further. For large data sets, the combination of improvements in the sequential algorithm and the parallelization results in overall speedup of a factor of 1000. Here, we demonstrate that i-ADHoRe is capable of processing much larger datasets than the current state-of-the-art tools. In particular, the complete Ensembl release 57 (36) data set that contains 49 eukaryotic genomes can be analyzed in 1 h (using 64 CPU cores), while producing highly accurate results.

MATERIALS AND METHODS

Data sets

The *Arabidopsis thaliana* and *V. vinifera* genomes together with gene family information were retrieved from PLAZA, an on-line plant comparative genomics resource (13) that provides gene families constructed with Tribe-MCL clustering (37) starting from an all-against-all BLAST (38) protein similarity search. The *E*-values and bit scores were saved, because these values are necessary for Cyntenator (39) and MCScan (40). The lengths for *Carica papaya* gene lists were also obtained via PLAZA. Animal genomes and families were downloaded from Ensembl (release 57) with the Ensembl Perl API (41). An all-against-all BLAST protein similarity search was done to obtain bit scores and *E*-values. An overview of all included species in both PLAZA and Ensembl is included in Supplementary Table S2.

Detection of collinearity

The initial steps of the algorithm (Supplementary Figure S1) are identical to i-ADHoRe 2.0; tandem duplicated genes are mapped to a single representative and for each pair of gene lists a *gene homology matrix* (GHM) is generated (Figure 1A). In this sparse matrix, pairs of homologous genes are represented as dots and as such collinear regions will appear as dense diagonals. Compared to the previous i-ADHoRe version, several major components of the algorithm were re-implemented for a better performance. First, the statistical validation of the clusters in the GHM was improved. To avoid inclusion of diagonals in the GHM generated merely by chance, the significance of each cluster is now estimated with a statistical model that takes into account the overall background density of the matrix. When multiple seeds (i.e. clusters with at least three homologous gene pairs that meet the initial criteria) were found, a correction for multiple hypothesis testing was done either with the Bonferroni or False Discovery Rate (FDR) (42,43) method.

Significant collinear regions found during this initial detection were converted into a *profile*, both collinear regions were aligned, i.e. homologous genes are placed in the same column adding gaps where necessary (Figure 1B). Like in previous versions of i-ADHoRe this alignment can be done by progressively applying the Needleman–Wunsch (pNW) algorithm or a greedy graph (GG) based alignment strategy. In version 3.0, a novel alignment algorithm (GG2), described in Fostier *et al.* (35), was implemented. Using this aligned profile, a new search is performed (Figure 1C), here a GHM is created

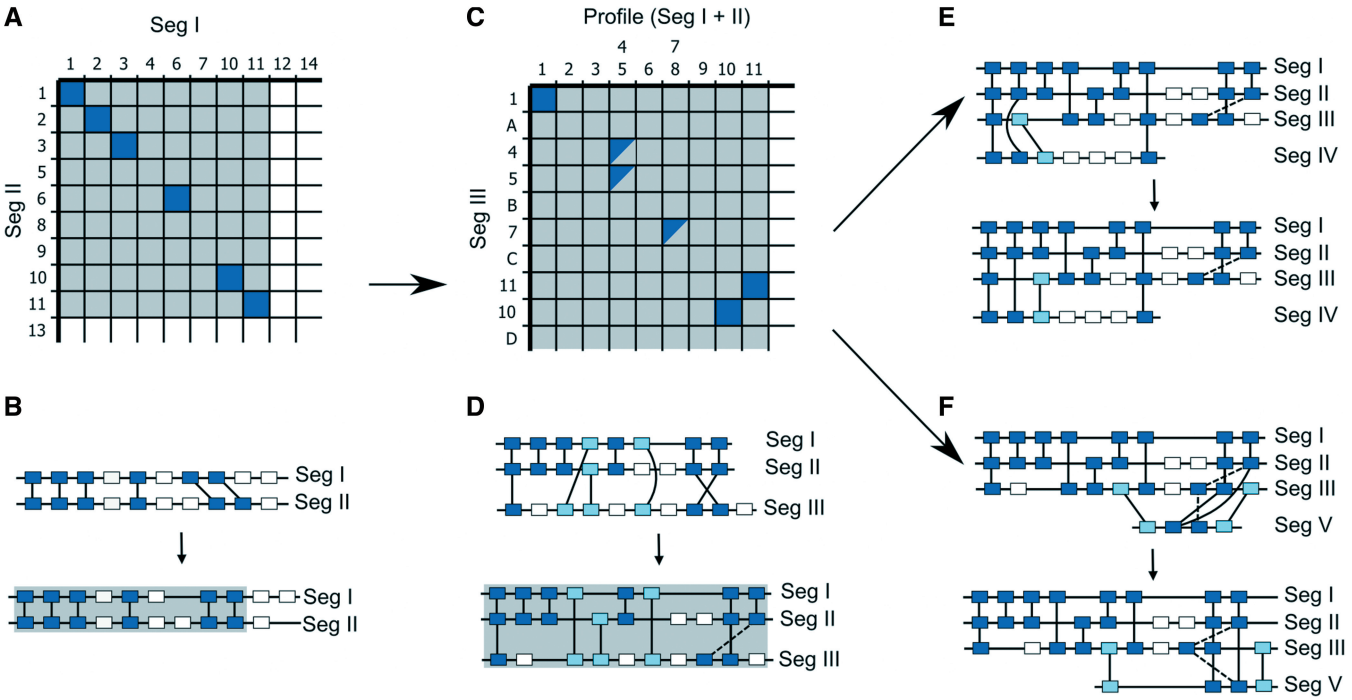

Figure 1. (**A**) GHM for the initial two segments (Seg I and Seg II). In a GHM, collinear regions will appear as dense diagonals. (**B**) Alignment of shared homologs between collinear regions; gaps are introduced to place as many homologous pairs in the same column as possible (35). The alignment (or 'profile') now contains the information of both segments. (**C**) Start of the iterative process, GHMs are now created with the profile and additional collinear regions can be found, e.g. Seg III. (**D**) Generation of a new profile. As long as additional segments can be found steps (C) and (D) are repeated. In this example (**E**) and (**F**) show how a single profile detects two additional segments that are mutually non-homologous (Seg IV and V), leading to a split in the detection process.

with the profile and all gene lists in the data set. Significant regions are added to a new profile and the profile search is repeated (Figure 1D). With a single profile, multiple segments can be found that are homologous to the profile but not necessarily to each other (Figure 1E and F). In this case, several profiles are generated and the detection algorithm continues detection with the longest profile first. Once no additional segments can be found the search continues with the next profile.

Additionally, the initial pairwise and profile searches can now be executed on a parallel computing platform (a multiprocessor/multicore systems or a computational cluster of networked computers). If N the number of gene lists provided as an input to i-ADHoRe, the $N(N+1)/2$ pairwise comparisons could be processed independently and, hence, distributed over different processes. The size of each gene list was taken into account to ensure a good load balance between the processes. At the end of this step, the detected collinear regions are communicated [using the Message Passing Inteface (MPI)] among the processes. Similarly, a single profile search can be parallelized by distributing the N gene lists among the different processes, again taking the size of the chromosomes into account. At the end of every profile search, the detected collinear regions were again communicated between the processes. However, due to the much smaller task granularity of one single profile search, a good load balancing was more difficult to achieve. Methods accompanying the novel synteny mode (detection of genomic homology purely based on shared gene content) can be found in the Supplementary Methods S1 and Table S3.

Empirical estimation of false positive rates

False positive (FP) rates were calculated with permutation tests in which 100 randomized data sets were compared with a real reference data set. Tandem duplicated genes (homologs within a window of 70 genes) were removed prior to shuffling the reference data set to generate a randomized version. This pre-processing step guaranteed a comparable density in the randomized run because breaking up tandem-clusters artificially increased the GHM background density. All genes had their original orientation replaced with a randomly assigned one. The lengths of the original gene lists were maintained during the randomization, but genes could be moved from one gene list to another. To estimate the performance with different settings, a permutation test was carried out for each of the desired settings, generating parameter landscapes for *Arabidopsis*, human and yeast, with various combinations of *q_value* and *gap_size* parameters. Settings that yielded the maximum amount of anchor points, while maintaining a FP rate near the selected cut-off value were considered as optimal (Supplementary Methods S2).

Comparison with MCScan and Cyntenator

BLASTP pairs for MCScan were filtered and only the best five hits in each species were retained (40). Because in MCScan first proteins are clustered to group homologous genes in gene families, this step was excluded when monitoring runtimes for the different tools. Cyntenator was also run with filtered BLASTP output, retaining only the top five hits for each species if their bit score was within 95% of the highest bit score [as described in Rödelsperger *et al.* (44)]. The gap and mismatch penalties were set to −0.3, the threshold to 2 and the filter to 1000. i-ADHoRe was run with a gap_size of 30 and cluster_gap of 35, while keeping the prob_cutoff on 0.01 and the *q*_value on 0.75. GG2 was used as the alignment algorithm and correction for multiple hypothesis testing was done with FDR. The minimal number of anchor points in a cluster was set to five.

Detection of highly conserved regions enriched for coexpressing/interacting gene pairs

Phylogenetic profiles, describing the number of homologous regions per species present in a *multiplicon*, a set of mutually collinear regions, were generated for all multiplicons in the output from the high-quality Ensembl subset. Multiplicons with one human and one bird (either chicken or zebra finch) segment and with conserved segments of at least five other mammals were selected. From these regions, the human segment was identified and the genes collinear with genes from other segments were stored. Expression data were derived from COXPRESdb version c3.1 (45) and highly expressed gene pairs were selected based on a mutual rank below or equal to 50. Experimentally characterized interacting protein pairs (41 088 binary interactions for 9142 human genes) were downloaded from IntAct (46). Using Ensembl's BioMart tool, a conversion table was generated to map all gene identifiers in these data sets to the Ensembl genes. For each selected multiplicon, the length of the human segment and number of human collinear genes were determined. Then, the number of coexpressed or interacting pairs was counted. When at least one human gene pair was found, the statistical significance was tested with a permutation test. Over 10 000 iterations, a random segment from the human genome (with the tandem duplicated genes removed) was sampled with the same length as the selected multiplicon. From the random region, an equal number of genes was randomly selected as collinear and, the number of coexpressed or protein–protein interaction pairs in this gene set was established. The number of iterations in which a number of pairs was equal or larger than that found in the real data set were counted and used to calculate a *P*-value for each multiplicon. All regions with a $P < 0.05$ were considered significant.

Evaluation of low quality genomes

To artificially reduce the quality of the *Arabidopsis* genome, the gene list length distribution of the papaya genome was used as a template to split the *Arabidopsis* gene lists in fragments resembling a draft assembly. i-ADHoRe was executed on both the *Arabidopsis* genome and the artificial low-quality version. The collinear fractions were measured by enabling the *write_stats* option in i-ADHoRe. Supplementary Methods S3 describes an additional study that further addresses this issue, using various vertebrate genomes.

RESULTS AND DISCUSSION

The i-ADHoRe 3.0 algorithm

The detection strategy of i-ADHoRe 3.0 is shown in Supplementary Figure S1 (25,47). First, tandem duplicated genes are mapped onto one single representative gene, because tandem clusters can hinder the detection of diagonals (see further). Next, for each pair of chromosomes or scaffolds, a so-called *gene homology matrix* (GHM) is generated. A GHM is a sparse matrix in which homologous gene pairs are marked by dots and collinear regions appear as 'diagonals'. For each detected diagonal, the statistical significance is evaluated (Figure 1A). Significant collinear regions are aligned into a *profile* (Figure 1B) that contains the combined gene content of the two collinear regions and can hence be used as a more sensitive probe to scan for additional collinear regions (Figure 1C and D). This step is iterated as long as new collinear regions are found and mutually homologous regions are grouped into a multiplicon. Even though the profile search requires an increased computational cost, it has proven its merits as a means to detect more degenerate genomic homology (12,25,48).

In order to deal with increasingly large data sets, various parts of the original i-ADHoRe code (47) have been replaced by equivalent algorithms with a reduced computational complexity. A first major improvement was the development of an efficient statistical model to estimate the significance of diagonals in the GHM, because the computational cost to calculate the exact *P*-value (49) increases exponentially with the number of gene pairs that shape the diagonal. The *Arabidopsis thaliana* data set was analyzed with different *P*-value thresholds and an empirical FP rate for each threshold was determined using permutation tests (Supplementary Figure S2). The combination of better heuristics and the implementation of a correction for multiple hypothesis testing (Bonferroni or FDR) resulted in a more realistic estimation of *P*-values and consequently improved the control of the FP rate compared to the previous statistical model. Benchmarks including other model organisms and the effects of using different parameter settings are reported in Supplementary Tables S4–S6.

In the iterative search procedure, additional collinear regions are identified and the corresponding profiles are updated in every step. Therefore, an accurate alignment algorithm is imperative for the sensitive discovery of more degenerate collinear regions (Figure 1C and D). Originally, i-ADHoRe relied on the progressive application of the pairwise Needleman–Wunsch (pNW) algorithm to align multiple homologous segments into profiles (47). Whereas with the Needleman–Wunsch

algorithm an optimal pairwise alignment of two segments can be obtained, its quality quickly degrades due to the propagation of erroneous decisions in early alignment steps when additional segments are added (50). To resolve this issue, a greedy, graph-based (GG) aligner had been introduced into i-ADHoRe 2.0 that converted the alignment problem into a cycle-canceling problem in a graph (25). Whereas this implementation provided a viable solution for the 'once a gap, always a gap' problem, it was unable to outperform the pNW aligner in terms of number of correctly aligned homologous genes. In i-ADHoRe 3.0, a novel greedy, graph-based aligner (GG2) was featured that, by means of maximum flow calculations in the graph, resolved efficiently unalignable sections in the graph (conflicts). Even though this graph-based method is computationally more intensive than the application of the pNW aligner, fast heuristics allow this algorithm to be efficiently used (35).

Finally, two practical issues arise when multiple genomes are compared: the processing time and the memory requirements. Whereas the runtime increases super-linearly with the size of the data set, i.e. faster than the number of genomes that are analyzed, the memory requirements are mainly determined by the number of homologous gene pairs. To limit the runtime and, hence, facilitate the analysis of large-scale data sets, the two most time-consuming parts of the algorithm were parallelized (Figure 1, green boxes): the initial all-to-all pairwise comparison (every gene list versus every gene list) and the iterative profile searches (one profile versus every gene list). The parallelization of the all-to-all pairwise step revealed that by using a dataset of 31 high-quality genomes (Supplementary Table S2) and 64 CPU cores, a 46-fold increase in speed (Supplementary Figure S3) was observed. Searching additional collinear regions in a gene list using a profile is more difficult to parallelize, because of more intense communication requirements between the subtasks and hence a larger communication overhead. Overall, the runtime for the complete algorithm was reduced 32-fold on 64 cores, corresponding to a parallel efficiency (relative reduction in runtime compared to one with one single core, over the number of cores used) of ~50%.

Evaluation of gene-based collinearity detection tools

When genomes with remnants of WGDs are dealt with or when highly diverged genomes are compared, gene loss and different types of rearrangements can interfere with the accurate detection of duplicated or homologous collinear regions (7,9). To the best of our knowledge, only Cyntenator (39), MCScan (40) and i-ADHoRe go beyond simple pairwise comparison and combine, via different approaches, information to find additional homologous regions. Cyntenator performs progressive pairwise combinations based on a user-defined species tree that strictly imposes the order in which genomes are compared. Only valid alignments including homologous regions from all species are retained to find collinearity with the next genome in line. Unlike the profile search of i-ADHoRe, in MCScan each chromosome is used as a reference and all pairwise collinear segments are mapped, followed by a multiple alignment procedure of homologous genes, inspired by the threaded blockset aligner (34). MCScan allows pairing regions that had initially not been detected based on their collinearity with the reference, a method referred to as 'transitive homology' (47). Unlike some tools (Supplementary Table S1), Cyntenator, MCScan and i-ADHoRe use ordered gene lists rather than the actual genome sequence. This level of abstraction allows for an efficient detection of collinearity. An additional advantage is that more diverged intergenic sequences do not interfere with the discovery of ancient collinearity or synteny.

To benchmark the application of a profile search versus transitive homology mapping of pairwise collinear segments, i-ADHoRe and MCScan were executed on the *Arabidopsis thaliana* data set to identify degenerated duplicated segments. Cyntenator was excluded from this experiment, because it does not allow detection of internal duplications. Figure 2 shows the number of genes present in regions with a certain level, indicating the total number of homologous segments. Although i-ADHoRe and MCScan use very different approaches, the number of genes in collinear regions was comparable (23 912 and 24 559, respectively), but the profile search enabled i-ADHoRe to group more genes in regions with level four (4499 versus 2669 genes), five (1223 versus 891) and six (1318 versus 340). This result implies that the more advanced profile search allows for a more sensitive detection of collinear regions compared to the progressive chaining in MCScan.

To evaluate the discovery of inter-species collinearity, the three tools were applied to analyze a small subset of the genomes available in Ensembl, namely human (*Homo sapiens*) (51), chimpanzee (*Pan troglodytes*) (52), mouse (*Mus musculus*) (53), chicken (*Gallus gallus*) (54) and pufferfish (*Tetraodon nigroviridis*) (55). For each gene, all overlapping homologous segments were retrieved and the highest number of species found in one single alignment (or multiplicon) was scored. In contrast to Cyntenator, MCScan and i-ADHoRe collapse tandem genes into one single representative and, therefore, reported fewer genes. The predefined species order applied by Cyntenator to compare genomes forms a major drawback for large-scale analyses including multiple species. For instance, a region that is collinear between human and mouse, but for which the homologous counterpart in the chimpanzee lineage was lost, will not be reported because only collinear regions from the first pairwise comparison (i.e. human and chimpanzee) are retained to identify additional collinearity in mouse. Therefore, a fair comparison was possible only for regions in which collinearity was conserved in all five species. Whereas using MCScan and Cyntenator, 416 and 498 genes were assigned to such regions, respectively, the profile search applied by i-ADHoRe allocated 3296 genes in multiplicons containing regions conserved in all five species (Supplementary Figure S4).

Fast algorithms that exhibit a favorable computational complexity are imperative to keep pace with the ever-increasing number of available genomes. Therefore,

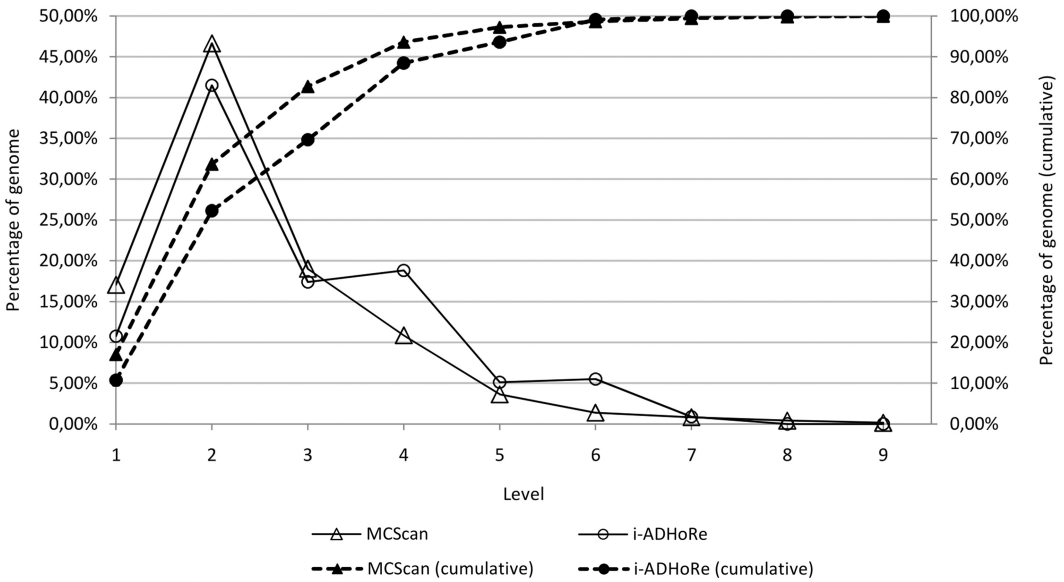

Figure 2. Distribution of the fraction of genes (*n*) found in sets of homologous genomic segments (multiplicons) with different levels (*m*) by MCScan and i-ADHoRe, respectively. Level 1 indicates the fraction of genes that was not found in any collinear region. The cumulative curve (i.e. the sum of all genes with the indicated level or lower) remains lower for i-ADHoRe, indicating that a larger fraction of the genome could be grouped into higher level multiplicons.

the runtime of all three programs was first monitored on the data set of the five species. i-ADHoRe, the only tool that takes advantage of a parallel environment, was executed using a single and eight threads, respectively, on a multicore machine. Because, MCScan first clusters proteins into gene families, a step not part of the actual collinearity detection algorithm, the program runtime was measured without this pre-processing step (Figure 3). Whereas Cyntenator required 6.25 h to analyze the five genomes, MCScan and i-ADHoRe were considerably faster, analyzing the data set in 19 and 14 min, respectively. When i-ADHoRe was run with eight cores, the runtime was reduced to only 3 min.

In a second experiment, the maximum number of genomes that could be analyzed was determined by processing data sets of gradually increased size (Figure 3). Only i-ADHoRe succeeded in analyzing the complete Ensembl data set covering 49 species (832 666 genes). Although Cyntenator could analyze up to 17 high-coverage genomes (39), the detection approach based on the strict usage of a guidance species tree posed a problem for data sets that include genomes sequenced at low coverage. As a result, inclusion of low-coverage or fractionated genomes into a large data set quickly eroded the amount of collinearity found, abruptly terminating the algorithm and leading to missing data when 10 or more genomes were included in the benchmark data set. For MCScan, the largest possible data set that could be analyzed in 48 h included 20 species (Figure 3); although within 168 h also 30 species could be covered, this duration however is impractically long for the efficient processing of extremely large data sets. In contrast, i-ADHoRe finished the full Ensembl data set covering 49 genomes within 42 h using a single CPU core. This runtime could be reduced to 6 h using the eight cores (88% efficiency). Finally, when using eight such machines (i.e. 64 cores in total) that are connected through a fast communication network (Infiniband), the runtime decreased to 40 min (50% efficiency; Supplementary Figure S3). An additional advantage of i-ADHoRe is that gene families rather than individual homologous gene pairs can be used to construct the GHM, whereas in both Cyntenator and MCScan, per query gene, a limit of five homologous genes in each other species (based on BLAST hits) is suggested. Furthermore, the usage of gene families is a more memory-efficient alternative than storage of all homologous gene pairs covering multiple genomes. Although for small data sets, i-ADHoRe utilizes more memory than MCScan and Cyntentor, the required memory scales linearly with the total number of genes and remains below that of MCScan once the data sets include 20 or more genomes (Figure 3).

Biological significance of ultra-conserved multispecies collinearity

Starting from 25 293 genomic scaffolds present in the Ensembl data set, 319 245 multiplicons were identified, some of which contained homologous regions from more than 20 species. The 'largest' multiplicon contained 33 segments from 22 species and included several homeobox Dlx proteins. Several HOX gene clusters including homeobox transcription factors were also found in a few high-level multiplicons (HOX D, level 28; HOX C and HOX D, level 22; HOX A and HOX D, level 25; HOX B and HOX D, level 20). This region is known to be highly conserved across species because these genes, involved in development of the body plan, require

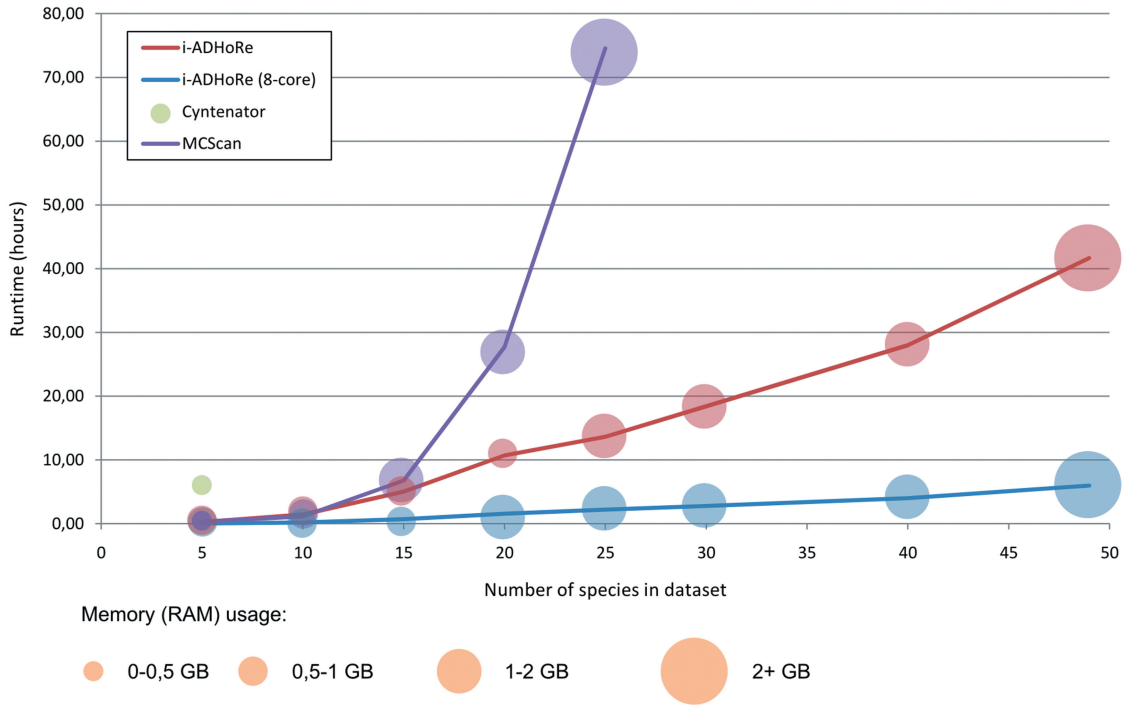

Figure 3. Runtime and memory usage comparison of Cyntenator (39), MCScan (40) and i-ADHoRe (this study). Each tool was run on subsets of the Ensembl data set each including a different number of species.

correct order to function (56). The HOX cluster was duplicated and retained during two rounds of WGD in the ancestor of the vertebrates over 450 Mya (21), and since then the HOX A, HOX B, HOX C and HOX D clusters diverged significantly (57).

Many genes coding for interacting proteins are robust against rearrangements (11) and clusters of coexpressed genes conserved between human and mouse have been reported (58). Given the large set of species, regions where gene order is strongly conserved over a large phylogenetic distance were delineated (see 'Materials and Methods' section). Next, we assessed whether genes in these strongly conserved regions showed significant functional clustering. Briefly, experimental protein–protein interaction data and coexpression information were used to determine whether a highly conserved region was significantly enriched for interacting proteins or genes showing coordinated expression profiles. Coexpression is frequently used as a strong indicator for functionally related genes ('guilt by association'). Also, interacting protein pairs are known to have a high chance to be involved in the same biological process (59). From the output of the high-quality subset, multiplicons with a strong conservation between either chicken or the songbird zebra finch (*Taeniopygia guttata*) (60), human, and at least five other mammals were extracted. Out of these 2863 multiplicons, 466 regions containing 2424 human genes, were significantly enriched ($P < 0.05$) for coexpressed pairs and/or gene pairs coding for interacting proteins (Figure 4). Mapping of these regions to a chromosome conservation plot depicting collinearity with all the 23 species included revealed that these regions are often among the most conserved in the genome (Supplementary Figure S5). A full list of conserved regions showing functional clustering, including the human genes within these regions and *P*-values for functional enrichment, can be found in Supplementary Table S7.

Significant enrichment of coexpressed and interacting protein pairs points toward an evolutionary constraint to conserve gene order in these regions. These results provide further evidence that gene order in vertebrates is non-random and might play a considerable role in regulation of gene expression. However, the precise mechanism of the observed coexpression remains an open question, because transcription factors, chromatin modifications (61), and long-range enhancers are likely candidates to play a role in this process.

Performance on low-coverage and fractionated genomes

Whereas new techniques greatly speed up the sequencing of novel genomes at a low cost, short-read lengths make it difficult to assemble reads into full chromosomes without genetic maps or a finishing phase (30,62,63). The same is true for genomes sequenced with traditional Sanger methods and low coverage (2×) (64). Consequently, these genomes are generally released as a collection of annotated scaffolds instead of being assembled into complete chromosomes or pseudomolecules. Draft-quality genomes are usually sequenced to get an overview of the overall gene content and, because most gaps occur in repeat regions, the majority of the genes are present despite the low coverage. However, for studies focusing on genome organization and evolution,

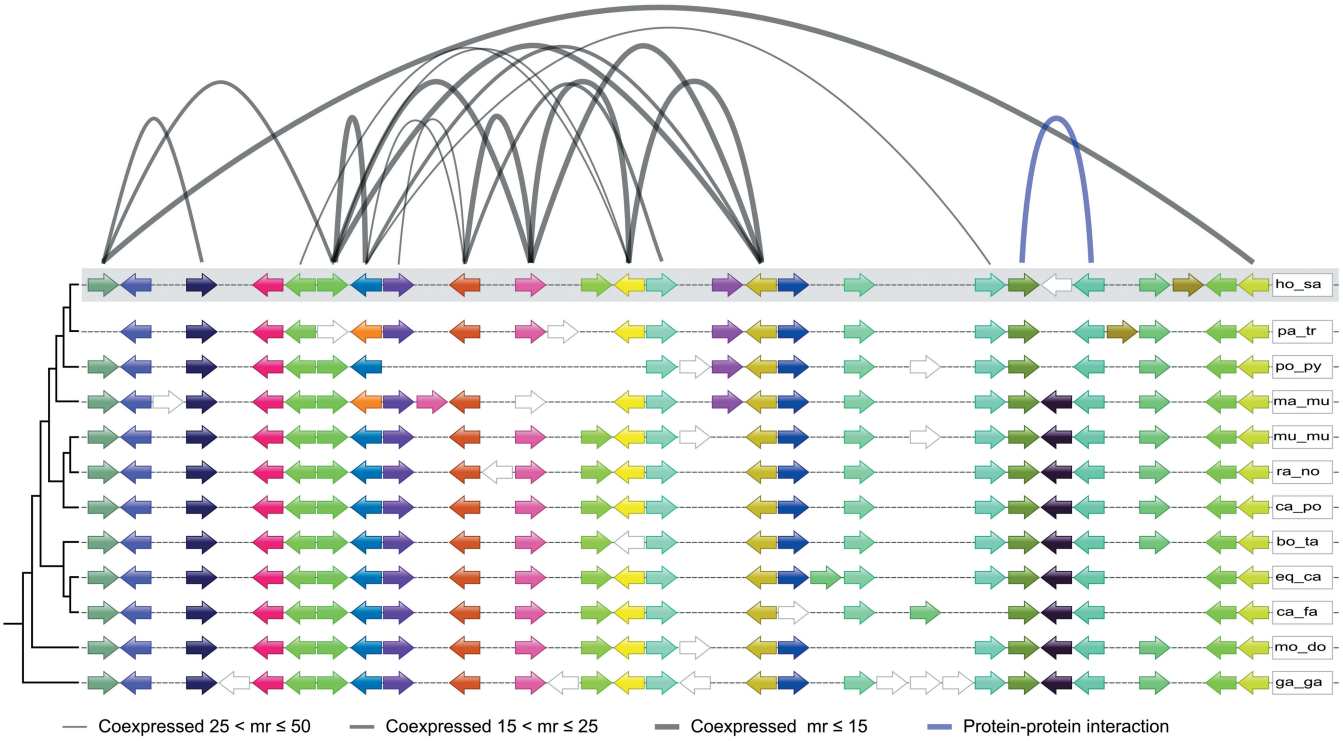

Figure 4. Gene order alignment of collinear regions conserved over a large phylogenetic distance (human–chicken). Species to which the segments belong are given on every line by the boxes on the right: *Homo sapiens* (ho_sa), *Pan troglodytes* (pa_tr), *Pongo pygmaeus* (po_py), *Macaca mulatta* (ma_mu), *Mus musculus* (mu_mu), *Rattus norvegicus* (ra_no), *Cavia porcellus* (ca_po), *Bos taurus* (bo_ta), *Equus caballus* (eq_ca), *Canis familiaris* (ca_fa), *Monodelphis domestica* (mo_do) and *Gallus gallus* (ga_ga). Arrows indicate coding genes and their orientation. Homologous genes are depicted in the same color. Coexpressed gene pairs are linked by black curved lines, of which the thickness of the line corresponds to the coexpression level (based on the mutual rank of the human genes in CoXPRESDB). Blue curved lines link pairs of genes coding for interacting proteins (in human). This region was found to be significantly enriched for coexpressed genes and, therefore, a biological constraint might cause gene order in this region to be retained.

inclusion of these low-quality genomes can become problematic (64). Consequently, we expect a genome sequence provided as a large set of unassembled scaffolds to interfere with the accurate detection of collinearity (Supplementary Table S8).

For instance, to estimate the impact of low-quality genomes on the detection of WGDs, the *Arabidopsis* genome was used as a reference and randomly cut into several artificial scaffolds with a length distribution comparable to the papaya (*Carica papaya*) genome (65) that is available as 4635 scaffolds (containing on an average six genes). The papaya genome was selected as a template because it is a draft version, sequenced up to 3× coverage, and without assembly in pseudomolecules. The *Arabidopsis* genome and the low-quality version were analyzed without any additional genomes. As expected, the number of genes that could be analyzed decreased because scaffolds with less than five genes were discarded (overall 17.47% gene loss). Whereas in the full genome 20 898 genes were found in duplicated regions (87.47%), this number decreased to only 10 091 (42.23% of all genes or 51.17% for the genes located only on scaffolds of sufficient length) in the low-quality version. Additionally, a drop in maximum level was observed: in the original genome up to seven homoeologous segments could be grouped whereas in the low-quality version the maximum level was five. However, including a genome reflecting the ancestral genome organization, like that of grapevine, can improve the number of genes found in collinear regions considerably. With grapevine included, 18% more *Arabidopsis* genes could be found in regions with level two to five (counting only *Arabidopsis* segments). Despite this increase in the number of genes found in duplicated regions, no more than five *Arabidopsis* segments grouped together after adding grapevine to the dataset. Because the maximum level is often used as a proxy for the number of large-scale duplication events (12), this result implied highly fragmented genomes organized in thousands of scaffolds are prone to misinterpretation when certain aspects of genome evolution are studied.

CONCLUSION

We show that the novel version of i-ADHoRe represents a major improvement over the current state-of-the-art algorithms and can be successfully applied to some of the largest data sets currently available.

As new sequencing initiatives such as the 1000 human genome project (66), the 1001 *Arabidopsis* genomes (67) and the 10 000 vertebrate genomes (68) will continue to generate many more genome sequences, the improved

scalability of i-ADHoRe is imperative to keep runtimes acceptable. The support for parallel computing platforms ensures that i-ADHoRe 3.0 will efficiently detect genomic homology and will be instrumental to unveil genome evolution in the different kingdoms of life.

AVAILABILITY

The i-ADHoRe 3.0 software package is available from http://bioinformatics.psb.ugent.be/software. Source code, documentation and example data sets are provided in the package.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–8, Supplementary Figures 1–5, Supplementary Methods 1–3 and Supplementary References [69–86].

ACKNOWLEDGEMENTS

The authors wish to thank Martine De Cock for help preparing the manuscript. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University.

FUNDING

Ghent University (Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks'); Interuniversity Attraction Poles Programme [IUAP P6/25] [initiated by the Belgian State, Science Policy Office (BioMaGNet)]; Agency for Innovation by Science and Technology in Flanders (predoctoral fellowship to S.P.); Postdoctoral Fellow of the Research Foundation-Flanders (to K.V.). Funding for Open Access Charge: Ghent University.

Conflict of interest statement. None declared.

REFERENCES

1. Garcia-Diaz,M. and Kunkel,T.A. (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem. Sci.*, **31**, 206–214.
2. Hurles,M. (2004) Gene duplication: the genomic trade in spare parts. *PLoS Biol.*, **2**, E206.
3. Comai,L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, **6**, 836–846.
4. Van de Peer,Y., Fawcett,J.A., Proost,S., Sterck,L. and Vandepoele,K. (2009) The flowering world: a tale of duplications. *Trends Plant Sci.*, **14**, 680–688.
5. Van de Peer,Y., Maere,S. and Meyer,A. (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, **10**, 725–732.
6. Passarge,E., Horsthemke,B. and Farber,R.A. (1999) Incorrect use of the term synteny. *Nat. Genet.*, **23**, 387.
7. Tang,H., Bowers,J.E., Wang,X., Ming,R., Alam,M. and Paterson,A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
8. Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.*, **2**, 333–341.
9. Van de Peer,Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.*, **5**, 752–763.
10. Stark,A., Lin,M.F., Kheradpour,P., Pedersen,J.S., Parts,L., Carlson,J.W., Crosby,M.A., Rasmussen,M.D., Roy,S., Deoras,A.N. *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.
11. Makino,T. and McLysaght,A. (2008) Interacting gene clusters and the evolution of the vertebrate immune system. *Mol. Biol. Evol.*, **25**, 1855–1862.
12. Simillion,C., Vandepoele,K., Van Montagu,M.C., Zabeau,M. and Van de Peer,Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **99**, 13627–13632.
13. Proost,S., Van Bel,M., Sterck,L., Billiau,K., Van Parys,T., Van de Peer,Y. and Vandepoele,K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.
14. Byrne,K.P. and Wolfe,K.H. (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, **175**, 1341–1350.
15. Thomas,B.C., Pedersen,B. and Freeling,M. (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**, 934–946.
16. Jiao,Y., Wickett,N.J., Ayyampalayam,S., Chanderbali,A.S., Landherr,L., Ralph,P.E., Tomsho,L.P., Hu,Y., Liang,H., Soltis,P.S. *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.
17. Fawcett,J.A., Maere,S. and Van de Peer,Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl Acad. Sci. USA*, **106**, 5737–5742.
18. Velasco,R., Zharkikh,A., Affourtit,J., Dhingra,A., Cestaro,A., Kalyanaraman,A., Fontana,P., Bhatnagar,S.K., Troggio,M., Pruss,D. *et al.* (2010) The genome of the domesticated apple (Malus x domestica Borkh.). *Nat. Genet.*, **42**, 833–839.
19. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
20. Vandepoele,K., De Vos,W., Taylor,J.S., Meyer,A. and Van de Peer,Y. (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. USA*, **101**, 1638–1643.
21. Dehal,P. and Boore,J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**, e314.
22. Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.
23. Kellis,M., Birren,B.W. and Lander,E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.
24. Scannell,D.R., Butler,G. and Wolfe,K.H. (2007) Yeast genome evolution–the origin of the species. *Yeast*, **24**, 929–942.
25. Simillion,C., Janssens,K., Sterck,L. and Van de Peer,Y. (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, **24**, 127–128.
26. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
27. Hu,T.T., Pattyn,P., Bakker,E.G., Cao,J., Cheng,J.F., Clark,R.M., Fahlgren,N., Fawcett,J.A., Grimwood,J., Gundlach,H. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.
28. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

29. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuveglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
30. Baliga,N.S., Bonneau,R., Facciotti,M.T., Pan,M., Glusman,G., Deutsch,E.W., Shannon,P., Chiu,Y., Weng,R.S., Gan,R.R. *et al.* (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. *Genome Res.*, **14**, 2221–2234.
31. Darling,A.C., Mau,B., Blattner,F.R. and Perna,N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
32. Dewey,C.N. (2007) Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol.*, **395**, 221–236.
33. Dewey,C.N. and Pachter,L. (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.*, **15(Spec No. 1)**, R51–R56.
34. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
35. Fostier,J., Proost,S., Dhoedt,B., Saeys,Y., Demeester,P., Van de Peer,Y. and Vandepoele,K. (2011) A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics*, **27**, 749–756.
36. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
37. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
38. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
39. Rodelsperger,C. and Dieterich,C. (2010) CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS One*, **5**, e8861.
40. Tang,H., Wang,X., Bowers,J.E., Ming,R., Alam,M. and Paterson,A.H. (2008) Unraveling ancient hexaploidy through multiplyaligned angiosperm gene maps. *Genome Res.*, **18**, 1944–54.
41. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
42. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Series B*, **57**, 289–300.
43. Dudoit,S. and van der Laan,M.J. (2008) *Multiple Testing Procedures with Applications to Genomics*. Springer, New York.
44. Rodelsperger,C. and Dieterich,C. (2008) Syntenator: multiple gene order alignments with a gene-specific scoring function. *Algorithms Mol. Biol.*, **3**, 14.
45. Obayashi,T., Hayashi,S., Shibaoka,M., Saeki,M., Ohta,H. and Kinoshita,K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, **36**, D77–D82.
46. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
47. Simillion,C., Vandepoele,K., Saeys,Y. and Van de Peer,Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.*, **14**, 1095–1106.
48. Vandepoele,K., Simillion,C. and Van de Peer,Y. (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.*, **18**, 606–608.
49. Durand,D. and Sankoff,D. (2003) Tests for gene clustering. *J. Comput. Biol.*, **10**, 453–482.
50. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
51. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
52. The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
53. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
54. International Chicken Genome Sequencing Consortium. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.
55. Jaillon,O., Aury,J.M., Brunet,F., Petit,J.L., Stange-Thomann,N., Mauceli,E., Bouneau,L., Fischer,C., Ozouf-Costaz,C., Bernot,A. *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.
56. Lewis,E.B. (1978) A gene complex controlling segmentation in Drosophila. *Nature*, **276**, 565–570.
57. Lemons,D. and McGinnis,W. (2006) Genomic evolution of Hox gene clusters. *Science*, **313**, 1918–1922.
58. Singer,G.A., Lloyd,A.T., Huminiecki,L.B. and Wolfe,K.H. (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.*, **22**, 767–775.
59. De Bodt,S., Proost,S., Vandepoele,K., Rouze,P. and Van de Peer,Y. (2009) Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, **10**, 288.
60. Warren,W.C., Clayton,D.F., Ellegren,H., Arnold,A.P., Hillier,L.W., Kunstner,A., Searle,S., White,S., Vilella,A.J., Fairley,S. *et al.* (2010) The genome of a songbird. *Nature*, **464**, 757–762.
61. Wu,C. (1997) Chromatin remodeling and the control of gene expression. *J. Biol. Chem.*, **272**, 28171–28174.
62. Harris,T.D., Buzby,P.R., Babcock,H., Beer,E., Bowers,J., Braslavsky,I., Causey,M., Colonell,J., Dimeo,J., Efcavitch,J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
63. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
64. Milinkovitch,M.C., Helaers,R., Depiereux,E., Tzika,A.C. and Gabaldon,T. (2010) 2x genomes–depth does matter. *Genome Biol.*, **11**, R16.
65. Ming,R., Hou,S., Feng,Y., Yu,Q., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.
66. Durbin,R.M., Abecasis,G.R., Altshuler,D.L., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
67. Weigel,D. and Mott,R. (2009) The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.*, **10**, 107.
68. Haussler,D., O'Brien,S.J., Ryder,O.A., Barker,F.K., Clamp,M., Crawford,A.J., Hanner,R., Hanotte,O., Johnson,W.E., McGuire,J.A. *et al.* (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, **100**, 659–674.
69. Proost,S., Pattyn,P., Gerats,T. and Van de Peer,Y. (2011) Journey through the past: 150 million years of plant genome evolution. *Plant J.*, **66**, 58–65.
70. Blanc,G., Hokamp,K. and Wolfe,K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.*, **13**, 137–144.
71. Locke,D.P., Hillier,L.W., Warren,W.C., Worley,K.C., Nazareth,L.V., Muzny,D.M., Yang,S.P., Wang,Z., Chinwalla,A.T., Minx,P. *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.
72. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L.,

Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
73. Vandepoele,K., Saeys,Y., Simillion,C., Raes,J. and Van De Peer,Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res.*, **12**, 1792–1801.
74. Hampson,S., McLysaght,A., Gaut,B. and Baldi,P. (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.*, **13**, 999–1010.
75. Hampson,S.E., Gaut,B.S. and Baldi,P. (2005) Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics*, **21**, 1339–1348.
76. Wang,X., Shi,X., Li,Z., Zhu,Q., Kong,L., Tang,W., Ge,S. and Luo,J. (2006) Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics*, **7**, 447.
77. Calabrese,P.P., Chakravarty,S. and Vision,T.J. (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19(Suppl. 1)**, i74–i80.
78. Pavesi,G., Mauri,G., Iannelli,F., Gissi,C. and Pesole,G. (2004) GeneSyn: a tool for detecting conserved gene order across genomes. *Bioinformatics*, **20**, 1472–1474.
79. Haas,B.J., Delcher,A.L., Wortman,J.R. and Salzberg,S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.
80. Hachiya,T., Osana,Y., Popendorf,K. and Sakakibara,Y. (2009) Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, **25**, 853–860.
81. Soderlund,C., Nelson,W., Shoemaker,A. and Paterson,A. (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.*, **16**, 1159–1168.
82. Soderlund,C., Bomhoff,M. and Nelson,W.M. (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.*, **39**, e68.
83. Cannon,S.B., Kozik,A., Chan,B., Michelmore,R. and Young,N.D. (2003) DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol.*, **4**, R68.
84. Sinha,A.U. and Meller,J. (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, **8**, 82.
85. Tang,H., Lyons,E., Pedersen,B., Schnable,J.C., Paterson,A.H. and Freeling,M. (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics*, **12**, 102.
86. Pham,S.K. and Pevzner,P.A. (2010) DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.